Local District Assessments: One Indicator for the Validation of Nebraska's Standards, Assessment, and Accountability System

Final Report
Prepared by:

Susan M. Brookhart, Ph.D.
DBA Brookhart Enterprises, LLC
2502 Gold Rush Avenue
Helena, MT 59601
(406) 442-8257
susanbrookhart@bresnan.net

For:

Pat Roschewski, Ph.D
Director of Statewide Assessment
Nebraska Department of Education
301 Centennial Mall South
Lincoln, Nebraska 68509

October, 2004
Revised December, 2004

Local District Assessments:
One Indicator for the Validation of Nebraska's
Standards, Assessment, and Accountability System

In order to monitor the progress of students in its public schools, the State of Nebraska uses a state assessment system based on standards of achievement for grades 4, 8, and 11. District reporting of student achievement is accomplished by locally-developed assessment systems that are reviewed for quality, thus having the effect of maintaining high standards for local assessment practices as well as supporting student achievement. Nebraska is known nationally for its position that local districts should be responsible for the assessment systems that monitor the progress of the students they teach (Roschewski, Gallagher, & Isernhagen, 2001).

The Nebraska STARS program is a unique state accountability system. School districts identify how they will measure and report student performance on content standards. They may select norm-referenced tests, develop criterion-referenced assessments, or use classroom assessments to measure state or state-approved content standards (Roschewski, 2004, p. 10).

The quality of district assessment *systems* has been evaluated annually since 2001 (Buckendahl, Plake, & Impara, 2004). Districts submitted portfolios in Reading in 2001 and 2003, and in Mathematics in 2002 and 2004. Six quality criteria (Plake, Impara, & Buckendahl, 2004) have been used to provide technical quality ratings for district assessment systems. This study is the first look at the quality of the local assessments used in those systems.

The Nebraska Department of Education prepared a draft evaluation and validation plan for its standards, assessment, and accountability system (Roschewski, February 2004). In the plan, each of five goals had several indicators for study. This study addressed the quality of local district assessments, an indicator for study listed under Goal One (Educators can appropriately and accurately assess and report student performance on content standards using local assessment systems).

The research questions for this study were the following:

1. Of what quality are the local assessments used in the Nebraska STARS program?
2. What proportion of them are of sufficient quality to accurately measure student performance?
3. If cases are found where quality is not deemed sufficient, what professional development and/or feedback to teachers might be required to raise the quality of assessment to an acceptable level?

Method

The general method used was threefold:

- Select a representative sample of local assessments in reading and mathematics.
- Work with Nebraska educators who already have training in assessment.

- Design a scoring workshop with three parts: (1) training on the use of rubrics for assessment quality, until a reliability criterion is reached; (2) evaluation of a sample of local assessments, using the rubrics; and (3) a short "debriefing" session.

Sample Assessments

In 2003 each Nebraska school district included, as an appendix to their Reading assessment portfolio, local assessments covering three Reading standards each at grades 4, 8, and 11. Similarly, in 2004 districts included local Mathematics assessments in their Math assessment portfolios. NDE randomly selected and randomly assigned content standards for the local assessment samples from among those not covered by NRT's. The local assessments themselves were not rated in the portfolio rating process.

A random sample, stratified by school district class (ESS/NDE, 2003), of 300 assessments was selected (50 each for 2 subjects at 3 grade levels) from the 2003 and 2004 district assessment portfolios. From the 300, 30 were randomly selected (5 each for 2 subjects at 3 grade levels); for these 30, two copies were included with the 300 assessments to be scored, thus embedding a double scoring study in the workshop. NDE staff pulled the assessments from portfolios, removed identifying information, and photocopied them.

Participants

Thirteen Nebraska educators, identified and invited by NDE, participated in the scoring workshop. Eight of these were graduates of the University of Nebraska-Lincoln assessment cohort program (Lukin, Bandalos, Eckhout, & Mickelson, 2004). All were experienced educators who had worked extensively with assessment. These educators will be using the assessment rubrics in professional development around the state in the coming year, so the scoring workshop functioned as training for them as well as a rating session.

The two-day scoring workshop was observed by the NDE Director of Statewide Assessment and three NDE curriculum coordinators (Reading, Math, and Social Studies), and by two UNL professors (including the coordinator of the STARS overall evaluation, of which this study is one part). The observers were present at the request of NDE, to keep the process open.

Instrument (Rubric) Development

Content Validity. Rubrics were developed with a dual purpose in mind. First, the rubrics were to be useful for the study described here. Second, the rubrics were to be sufficiently user-friendly and in line with other Nebraska STARS materials to be useful in district professional development. The goal was to study the quality of local assessments currently used in STARS, and then to use both the results and the rubrics to improve the quality of local assessments. The rubrics for this study were developed by analyzing and comparing several existing sets of rubrics for judging the quality of classroom assessments: the Alternative Assessment Checklist (NWREL, 1998), the Classroom Assessment Quality Rubrics (Arter & Busick, 2001), the CRESST Language Arts Assignment Rubric (Matsumura *et al*., 2002), and rubrics developed for

assessing student teachers' assessments (McConney & Ayres, 1998). Current conceptions of reliability and validity for classroom assessments (Brookhart, 2004) were considered.

Draft rubrics were constructed and revised in consultation with the NDE Director of Statewide Assessment in order to be consistent with other NDE assessment communications (NDE, 2003). The resulting rubrics had five traits: Alignment, Sufficiency, Clarity, Appropriateness, and Scoring Procedures. A copy of the rubrics used in the study is included in Appendix A, Scoring Workshop Session Plans. It is important to note that there are at least two additional criteria for the quality of local assessments that could not be measured for this study: amount of student involvement and integration with instruction (opportunity to learn/practice). Investigating these would have required information outside of the scope of this study.

For use in the workshop, the rubrics were printed on one page. Space was provided for recording codes for the assessment ID number, district, and standard. The resulting one-page instrument was used (already filled in) for example assessments, and for training and general rating (one sheet per rater/assessment). Appendix A presents directions and instruments for replicating the scoring workshop.

Reliability. Three levels of performance were described for each trait: low (1-2 points), medium (3-4 points), and high (5-6 points). Full points (2, 4, or 6) were awarded if the criteria for that level were met in full, the lesser point amount (1, 3, or 5) if the description applied but was not met in full.

The first day of the scoring workshop was a training day. First, four example assessments were discussed, and then five training assessments were scored independently. Rater agreement overall (13 raters rating 5 training assessments on 5 criteria) was 86% exact agreement on category (low/medium/high) and 66% exact agreement on point value (1-6). The second day, raters scored the remaining assessments independently.

Results from the 30 double-scored papers indicated that Math ratings remained reliable but Reading ratings did not. Intraclass correlations for total score (for a single rater) were .59 for Math and -.13 for Reading. Where subsets of double-scored papers were scored by the same two raters (3 Math subsets and 2 Reading subsets, a total of 20 of the 30 papers), generalizability analyses were possible (Crick & Brennan, 2001). Generalizability values for one rater ranged from .78 to .90 (mdn=.81) for relative decisions and .65 to .88 (mdn=.81) for absolute decisions in Math and were .00 for Reading. More detailed results of the double scoring study are presented in Appendix B.

Differences in reliability for Math and Reading in the double-scoring – after all raters had reached a reliability criterion in general training – suggest that the reliability issue is more about applying the criteria to the Reading assessments than about the Reading raters.

Results

Research Questions 1 and 2: The Quality of Local District Assessments in the Nebraska STARS

1. Of what quality are the local assessments used in the Nebraska STARS program?

The quality of local assessments used in the Nebraska STARS program was good overall. For three of the criteria (Alignment, mean=5.24, sd=1.17; Clarity, mean=5.32, sd=1.19; and Appropriateness, mean=5.28, sd=.99), the majority of local assessments received a score of 5 or 6 (in the top rubric category). Ratings for Scoring Procedures were lower (mean = 3.16, sd=1.72), mostly because scoring procedures were not provided for many of the assessments. One of the participants described this as an "easy fix," and pointed out that the rubrics would be an effective way of communicating with districts the importance of including scoring procedures in local assessments. Scores for Sufficiency (mean = 4.59, sd=1.20) also indicated an area for improvement. Many of the assessments did not have enough information available at all performance levels (beginning, progressing, proficient, and advanced) to reliably classify students at all performance categories.

Table 1 below presents overall descriptive statistics for the assessments.

Table 1. Mean Assessment Ratings, Overall and by Grade

Grade		Alignment	Sufficiency	Clarity	Appropriate-ness	Scoring Procedures
4	Mean	5.31	4.67	5.41	5.19	3.39
	s.d.	1.18	1.21	1.17	1.05	1.74
	n	99	99	99	99	99
8	Mean	5.28	4.57	5.28	5.16	3.18
	s.d.	.96	1.16	1.28	1.00	1.81
	n	96	96	94	96	96
11	Mean	5.11	4.52	5.27	5.49	2.90
	s.d.	1.32	1.25	1.12	.90	1.60
	n	98	98	98	98	98
Total	Mean	5.24	4.59	5.32	5.28	3.16
	s.d.	1.17	1.20	1.19	.99	1.72
	n	293	293	291	293	293

Tables 2 and 3 below present statistics disaggregated by subject (Math, Reading) and grade level (4, 8, 11).

Table 2. Mean Mathematics Assessment Ratings

Grade		Alignment	Sufficiency	Clarity	Appropriate-ness	Scoring Procedures

4	Mean	5.67	4.96	5.59	5.43	3.14
	s.d.	.69	.84	.73	.76	1.79
	n	49	49	49	49	49
8	Mean	5.42	4.58	5.84	5.24	3.08
	s.d.	.88	.81	.43	.92	1.72
	n	50	50	49	50	50
11	Mean	5.33	4.65	5.44	5.48	2.37
	s.d.	1.17	1.04	.87	.97	1.32
	n	48	48	48	48	48
Total	Mean	5.48	4.73	5.62	5.38	2.87
	s.d.	.94	.91	.72	.89	1.65
	n	147	147	146	147	147

Table 3. Mean Reading Assessment Ratings

Grade		**Alignment**	**Sufficiency**	**Clarity**	**Appropriate-ness**	**Scoring Procedures**
4	Mean	4.96	4.38	5.24	4.96	3.64
	s.d.	1.44	1.44	1.47	1.23	1.66
	n	50	50	50	50	50
8	Mean	5.13	4.57	4.67	5.07	3.28
	s.d.	1.02	1.46	1.60	1.08	1.92
	n	46	46	45	46	46
11	Mean	4.90	4.40	5.10	5.50	3.40
	s.d.	1.43	1.41	1.30	.84	1.69
	n	50	50	50	50	50
Total	Mean	4.99	4.45	5.01	5.18	3.45
	s.d.	1.32	1.43	1.46	1.08	1.75
	n	146	146	145	146	146

A MANOVA using the sample of 291 assessments with data on all criteria tested the statistical significance of the differences among groups. Mathematics assessments were of higher quality than Reading assessments in Alignment, Sufficiency, and Clarity. There was no significant difference between Math and Reading for Appropriateness. Reading assessments scored higher in Scoring Procedures than did Math. The effect for subject was significant (Wilks' lambda for the subject effect = .85, $F(5,281)=9.91$, $p=.00$). The effect for grade was not significant (Wilks' lambda for the grade effect = .94, $F(10,562)=1.80$, $p=.06$), and neither was the interaction of subject and grade (Wilks' lambda for the interaction effect = .94, $F(10,562)=1.77$, $p=.06$). Univariate follow-up results of the MANOVA are presented in Table 4 and add further detail. In summary, the only observed differences in the means tables above that were significant were the following:

- differences between Math and Reading on Alignment, Sufficiency, Clarity, and Scoring Procedures

- in Math, grade 8 assessments were the most clear; in Reading, grade 8 assessments were the least clear (the interaction effect for Clarity).

Table 4. Univariate Follow-up of MANOVA by Subject and Grade			
Criterion	**Subject Effect (Math, Reading)**	**Grade Effect (4, 8, 11)**	**Interaction (effect of grade different for each subject)**
Alignment	**Yes** F(1,285)=12.41, p=.00	**No** F(2,285)= .89, p=.41	**No** F(2,285)= .92, p=.40
Sufficiency	**Yes** F(1,285)=3.94, p=.05	**No** F(2,285)= .37, p=.69	**No** F(2,285)= 1.35, p=.26
Clarity	**Yes** F(1,285)=21.50, p=.00	**No** F(2,285)= .61, p=.54	**Yes** F(2,285)= 4.17, p=.02
Appropriateness	**No** F(1,285)=3.57, p=.06	**No*** F(2,285)= 3.08, p=.05	**No** F(2,285)= 1.57, p=.21
Scoring Procedures	**Yes** F(1,285)=8.26, p=.00	**No** F(2,285)= 2.25, p=.11	**No** F(2,285)= 1.50, p=.22

*An apparent grade effect for Appropriateness did not show any group differences in a Scheffe post-hoc test, consistent with the multivariate test results.

2. What proportion of assessments are of sufficient quality to accurately measure student performance?

Tables 5 through 9 below present the percent of all assessments at each level for Alignment, Sufficiency, Clarity, Appropriateness, and Scoring Procedures. Appendix C presents the number and percent of assessments at the various quality levels by grade and subject. The disaggregated data show patterns similar to the overall data.

Overall, approximately 74%, or 3 out of 4, assessments were comprised of items or tasks reflecting a match to the standard they were intended to measure. Approximately 54% had an appropriate number of score points at all 4 performance levels (Beginning, Progressing, Proficient, and Advanced). Approximately 80% of assessments had tasks and directions that were clear to students. Approximately 79% of assessments were judged appropriate (fair for all students, and of an appropriate level and length for intended grade level). Approximately 26% specified clear scoring procedures consistent with the task; 59% had scoring procedures that were either unclear or not provided. Assuming many districts did not provide scoring procedures when asked for their assessments – interpreting that request to mean providing a copy of the student instrument – this figure underestimates the quality of the scoring procedures. Therefore, the majority of local assessments were of sufficient quality on characteristics that go to validity (alignment with standards, clarity to students, appropriateness of content). More work needs to be done to raise the quality of local assessments on aspects related to reliability (sufficiency of information and scoring procedures).

Table 5. Alignment – Number and Percent of Assessments at Each Level of Quality

(Total = 293 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Assessment items/tasks reflect a match to the appropriate standard(s).	6	183	62.5
	5	34	11.6
Some of the assessment items/tasks reflect a match to the appropriate standard(s).	4	54	18.4
	3	10	3.4
Few of the assessment items/tasks reflect a match to the appropriate standard.	2	8	2.7
	1	4	1.4

Table 6. Sufficiency – Number and Percent of Assessments at Each Level of Quality (Total = 293 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
The essence of the standard is represented by an appropriate number of score points at all four performance levels (required by 2006-07).	6	77	26.3
	5	82	28.0
The essence of the standard is represented by an inadequate number of score points and only at certain performance levels.	4	101	34.5
	3	6	2.0
The essence of the standard is not represented, is ignored, or is poorly sampled.	2	23	7.8
	1	4	1.4

Table 7. Clarity – Number and Percent of Assessments at Each Level of Quality (Total = 291 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
The assessment tasks and the directions are clear, complete, and unambiguous.	6	197	67.7
	5	35	12.0
Some of the assessment tasks or the directions are clear, complete, or unambiguous.	4	36	12.4
	3	3	1.0
Few of the assessment tasks and/or directions are clear, complete, or unambiguous.	2	18	6.2
	1	2	0.7

Table 8. Appropriateness – Number and Percent of Assessments at Each Level of Quality (Total = 293 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Assessment tasks are fair, at the appropriate cognitive level, and of an appropriate length.	6	166	56.7
	5	64	21.8
Some of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length.	4	50	17.1
	3	6	2.0
Few of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length.	2	6	2.0
	1	1	0.3

Table 9. Scoring Procedures – Number and Percent of Assessments at Each Level of Quality (Total = 293 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Scoring procedures/ rubrics are consistent with the assessment task and can be applied clearly and consistently.	6	58	19.8
	5	17	5.8
Scoring procedures/ rubrics are provided, but require judgments that would be hard to make fairly and consistently.	4	36	12.3
	3	9	3.1
Scoring procedures/rubrics are inadequate or not provided.	2	148	50.5
	1	25	8.5

Correlation analysis investigated the question of whether the quality of local assessments is related to the quality of the assessment system in a district. Most districts sampled were represented by 2 or 3 assessments. The median score for assessments in each district at each grade was matched with that district/grade's assessment system rating (1-5, unacceptable through exemplary, based on the 2003 Reading and 2004 Math portfolio ratings). Spearman's rho, a correlation statistic appropriate for ordinal scale data, was used. Table 10 below presents the results.

Table 10. Correlations between District Assessment Portfolio Ratings and Median Ratings of Assessments Sampled from Those Portfolios		
Assessment Quality Criterion	**Reading (2003) n = 53**	**Math (2004) n = 51**
Alignment	.15	-.26
Sufficiency	.05	.29*
Clarity	-.08	-.04
Appropriateness	-.09	.12
Scoring Procedures	.15	.04

* $p < .05$

The quality of local assessment systems, as measured by portfolio ratings, was not related to the quality of local assessment instruments. One correlation reached statistical significance (between Sufficiency and assessment quality for Math portfolios); however, at the .05 level one in 20 (or about .5 in 10, as here) tests would be expected to be significant by chance. It may be that there really is a relationship between Sufficiency and the quality of the assessment portfolios in Math, or it may be chance. None of the other relationships reached significance. The positive and negative correlations observed are just chance variation around zero.

There are many possible reasons for the lack of relationship between the quality of the district's assessment system, as reflected in the portfolio rating, and the quality of the local assessments. Lack of variability is clearly one of the reasons. In Reading, 89% of the district assessment system ratings were Very Good (40%) or Exemplary (49%). In Math, 96% of the district assessment system ratings were Very Good (16%) or Exemplary (80%). As Tables 5 through 9 showed, most of the ratings of individual assessments are also clustered at the top, although not quite as dramatically as the assessment system ratings, and the medians for a given district/grade formed similarly skewed distributions.

In addition to this statistical reason for lack of relationship between district assessment systems and assessments, there are several other possible explanations. It may be that the district assessment system rating was too aggregated a measure (that, for example, portfolio ratings about match to standard would have correlated with the assessments' alignment ratings, when the summary rating did not). It may be that those who wrote the assessments were fairly good at it, irrespective of how well the district system to review and approve assessments functioned. It may be that portfolio ratings are partly a measure of how good the district is at expressing what they did to assure assessment quality. While any of these reasons may be true, the lack of variability of the portfolio ratings, especially in Math, was so dramatic as to make any of the other explanations pale in comparison.

This analysis helps focus the professional development efforts (below). The fact that the quality of the local assessment instruments is not related to quality of the district assessment plans suggests a need for across-the-board professional development in classroom assessment. If the quality of local assessment instruments had been poorer in districts that needed help with their assessment plans as well, that would have suggested a need for professional development in classroom assessment and in district assessment planning targeted to particular districts.

Research Question 3: Professional Development Needs

The third question – If cases are found where quality is not deemed sufficient, what professional development and/or feedback to teachers might be required to raise the quality of assessment to an acceptable level? – was addressed in two ways. First, the rating patterns on the various rubric criteria gave a general indication of the kinds of improvements needed. Teachers are understandably quite good at matching assessments to standards and to instruction. "Understandably," because this is the main focus of the work teachers plan and ask students to do daily. This study's results for Sufficiency and Scoring Procedures indicated that teachers

would benefit from professional development on the ways in which scoring operates to turn student performance on that work into measurement.

Second, teachers responded to questions about professional development in the debriefing session. Appendix D presents verbatim comments the scorers wrote on their debriefing questionnaires and the facilitator's notes from the debriefing session. Professional development themes from these two sources included:

- Looking at ***sample assessments*** and discussing their quality is a powerful professional development tool.
- The ***rubric*** itself is valuable, making quality criteria explicit.
- There was support and approval for state involvement in ***improving assessment literacy***.

These educators were describing how having criteria (expressed in the rubrics) and examples (sample assessments) leads to learning (improved assessment literacy). There was enthusiastic support for the idea that discussions over example assessments would increase assessment literacy.

Beyond raising the general level of assessment literacy among Nebraska educators, participants called for some specific work regarding:

- Sufficiency – what it is and how to represent performance at four levels. Many of the assessments were actually mastery/nonmastery tests, with very little coverage at the beginning or advanced levels.
- Scoring procedures – what they are, why they are necessary, and how to write them up well.
- Alignment – teachers' interpretations of the standards need to "capture the essence" of the standard, not slavishly represent all the example indicators. Those are optional and are meant to be examples (and these participants did not think they were necessarily always the best examples).
- Content knowledge – in some cases assessments did not represent a deep knowledge base on the part of the assessment writers.
- Formatting – the design and presentation of questions, answer spaces, and so on.
- Guidance on what constitutes "appropriateness" (vocabulary, length, appropriateness of content to expected level of understanding and performance) over a range of kinds of standards (broad or narrow, written at different cognitive levels).

Discussion

The Nebraska STARS program has received national attention since its inception (Roschewski, Gallagher, & Isernhagen, 2001). Local assessment has a prominent role in state accountability reporting (STARS means School-based Teacher-led Assessment and Reporting System). Until now, districts' assessment systems have been evaluated (Buckendahl, Plake, & Impara, 2004), and the quality of individual local assessments has been assumed. This study addressed that assumption with empirical evidence. The majority of local assessments were of sufficient alignment, clarity, and appropriateness to warrant attention to their results. The results for sufficiency, coupled with participants' observations about lack of sufficiency being more of an issue for low and high performers, suggest that the quality of assessments may be of sufficient

quality at the middle of the range of student performance. That is, Nebraska can be confident about the accuracy of estimation of the percent of students above and below the progressing/proficient cutpoint; there is less confidence about the precise percentage of beginning and advanced students.

Even with a need for increased reliability (similar to the development of reliability in the assessment system portfolio process; Buckendahl, Plake, & Impara, 2004), this study has identified two areas that can with some confidence be shared with school districts as areas for improvement. Attention should be given to developing explicit scoring procedures for local assessments and to sufficient coverage at all four performance levels: beginning, progressing, proficient, and advanced. Suggestions for professional development from participants included attention to the following: sufficiency, scoring procedures, alignment (especially interpretation of standards), depth of content knowledge, formatting, and appropriateness.

Since this study was designed, another set of criteria and rubrics for judging alignment of state tests to standards, those used by Achieve, Inc., has been published (Resnick, Rothman, Slattery, & Vranek, 2003-04). Achieve's rubrics would not have been suitable for this Nebraska study because the Achieve ratings included the possibility of judging the *standards* as not specific enough, and that was not in the purview of this study. Also, it assumed multiple standards might be covered by a test, which is different from the way Nebraska districts sent in sample assessments by individual standard. However, it is important to note that the criteria used are parallel with the criteria used in the Nebraska study. Regarding results, states (5 of Achieve's clients) were found to have items that mapped well for content and performance, but did not fare well on balance and range. Regarding level of challenge, they found that the most challenging standards were those that were poorly represented on tests. These results parallel the findings of this study of Nebraska local assessments, where sufficiency of information at all performance levels was the most serious lack. Thus it seems that Nebraska has identified an area for improvement that it shares with other states around the country.

Suggestions for Rubric Revisions

The development of rubrics that will be useful for further assessment literacy development in the state is important. Overall, participants loved the rubrics, and think they will – even as is – help focus teacher understanding of high quality assessments and their efforts at writing them. Suggestions for improvement follow. For clarity in the discussion below, "teachers" will mean local educators who developed the local assessments (even though ESU staff were also involved), and "participants" will mean the panel of Nebraska educators who rated the local assessments in this study.

Alignment – Some of the teachers' assessments interpreted "match" to mean any match, including items that covered only a small subset of what a standard intended. Participants liked the language "accurate reflection of the standard" to mean a "match" to the intent of the standard. Or "match and represent well" or something like that. Participants found the concept of the "essence of the standard" helpful when thinking about alignment (and the author think the

"essence" goes here, too), although the use of that phrase is not necessary if one uses "accurate reflection" or "represent well."

Sufficiency – Here, the "essence of the standard" phrase led some of the participants to confuse sufficiency judgments with alignment judgments a bit. A suggestion for the 5-6 box would read, "There are an appropriate number of score points at all four performance levels," and for the 3-4 box, "There are an inadequate number of score points." If participants could circle the first part of the top box (appropriate number of score points), but not the second (all four levels), they were scoring that assessment a 5. The middle box, that says at only certain performance levels, became redundant to that.

Clarity – The clarity rubric was clear (pun intended). Participants were asked to consider the clarity rubric to mean "clarity to students": "If I were a kid, would I know what to do?"

Appropriateness – Dr. Roschewski and the author have already talked about replacing "cognitive level" with "grade level." Perhaps additional revision might be helpful, in case someone raises the issue of students not "on" grade level; however, grade level is much better than cognitive level here.

Scoring Procedures – Participants did not have any suggestions for rubric revision. An attempt to say scoring procedures could be counted as "provided" if all that was needed was an answer key failed. Therefore, "not provided" ended up strictly interpreted. However, for coaching purposes in a district, if all that was needed was an answer key, participants noted that that would be an easy fix. The author was surprised that participants did not see a clearer difference between right/wrong scoring of one-point answers and some other kinds of answer keys (e.g., right/wrong answers but multipoint scoring, or simple constructed response answers where not all students would write exactly the same thing). This is further evidence that scoring (in the sense of what are "points" and how does one go about allocating those) might be a useful topic for professional development in the future somewhere.

For use around the state, the participants also suggested a "Comments" space at the bottom for written feedback. This is a good idea. However, the rubric should not be crowded. Part of what it has going for it is its visual simplicity, which will aid users in committing these principles of good practice into their memories and into their own practice.

Follow-up Possibilities

There are at least three follow-up possibilities that stem from this study.

1. The state is already planning assessment quality workshops as part of the 2004-2005 "Chats" around the state. Examining sample assessments and working with the rubrics is the main follow-up to this study. It is recommended that some data be gathered about the effectiveness of these workshops.

2. In-depth training on the nature of "scoring" as a means of translating student performance on assessment items or tasks into measurement on a continuum of levels of student achievement regarding a standard would be helpful. Understanding the concept of "how the numbers work" to do that and understanding how difficulty levels of items and tasks map the achievement continuum would help teachers write assessments with sufficiency of coverage. Judging from the assessments reviewed, teacher understanding of scaling and scoring issues demonstrated a weaker conceptual base than teacher understanding of the standards. Basic understanding of scoring concepts should also raise assessment literacy and assessment quality in general.

3. Dissemination of the finding that local assessments, even at first cut, are of reasonable quality would be very helpful for the future of the Nebraska STARS program. The author and Dr. Roschewski have begun this process with two conference proposals. The 2005 Nebraska Assessment Conference will also help disseminate results. STARS has bet on the quality of local assessments, and it seems that was a safe bet. Dissemination of the rubric itself would also be helpful for other states that want to raise assessment literacy. Dissemination of the specific recommendations for professional development, especially as they match what other states, evaluators and researchers are finding in other locations, will provide information to focus professional development that might advance the quality of local assessments nationally.

References

Arter, J. A., & Busick, K. U. (2001). *Practice with student-involved classroom assessment*. Portland, OR: Assessment Training Institute.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice, 22*(4), 5-12.

Buckendahl, C. W., Plake, B. S., & Impara, J. C. (2004). A strategy for evaluating district developed assessments for state accountability. *Educational Measurement: Issues and Practice, 23*(2), 17-25.

Crick, J. E., & Brennan, R. L. (2001, January). GENOVA (version 3.1). Available http://www.uiowa.edu/~itp/pages/SWGENOVA.SHTML

Education Support Services, Nebraska Department of Education. (2003, March). *Statistics and facts about Nebraska schools*. Available: http://ess.nde.state.ne.us/DataCenter/PDF/0203/0203StatsFacts.pdf

Lukin, L. E., Bandalos, D. L., Eckhout, T. J., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice, 23*(2), 26-32.

Matsumura, L. C., Garnier, J., Pascal, J., & Valdes, R. (2002). Measuring instructional quality in accountability systems: Classroom assignments and student achievement. *Educational Assessment, 8*, 207-229.

McConney, A., & Ayres, R. R. (1998). Assessing student teachers' assessments. *Journal of Teacher Education, 49*, 140-150.

Nebraska Department of Education. (2003, March). *The school district assessment portfolio instructions and suggestions.*

Northwest Regional Educational Laboratory. (1998). *Toolkit 98*. Portland, OR: Author.

Plake, B. S., Impara, J. C., & Buckendahl, C. W. (2004). Technical quality criteria for evaluating district assessment portfolios used in the Nebraska STARS. *Educational Measurement: Issues and Practice, 23*(2), 12-16.

Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2003-2004). Benchmarking and alignment of standards and testing. *Educational Assessment, 9*, 1-27.

Roschewski, P. (2004). History and background of Nebraska's School-based Teacher-led Assessment and Reporting System (STARS). *Educational Measurement: Issues and Practice, 23*(2), 9-11.

Roschewski, P. (2004, February). Validation of Nebraska's standards, assessment, and accountability system. Administrative memo, Nebraska Department of Education.

Roschewski, P., Gallagher, C., & Isernhagen, J. (2001). Nebraskans reach for the STARS. *Phi Delta Kappan, 82*, 611-615.

Local District Assessments:
One Indicator for the Validation of Nebraska's
Standards, Assessment, and Accountability System

Appendix A

Scoring Workshop Session Plans

Scoring Workshop Session Plans
As Used July 13-14, 2004
Could be Adapted for State Workshops

Preparing for the training session

Identify the teachers and facilitators. Issue invitations.
Secure the location and setup (tables and chairs, refreshments, etc.). Setup should include tables with plenty of work space.
Make scoring packets. Include (copies are in this appendix, except for state standards):

- Agenda
- Purpose of Study
- Confidentiality agreement
- Rubrics
- Debriefing Questionnaire
- Copies of Nebraska Reading and Math standards
- "Do's and Don'ts for Item Writing"

Assemble additional materials

- Exemplar Assessments (at least 4) to Illustrate Rubric Levels
- Training Assessments (5-8)
    - Examples and training assessments were selected by the facilitator for the July 13-14 workshop, from the sample of local assessments for this study. For statewide workshops, local assessments might be selected from participating districts.
- Blank rubric forms (one for each assessment to be rated)
- Assessments to be reviewed and rated
- Pens/pencils
- Calculator
- Training paper score sheets

Conducting the workshop

Make introductions around the room. Review the "purpose" page. Describe where the assessments came from (for July 13-14, from a random sample of assessments in Reading and Math portfolios; for state workshops, from local districts or ESU's) and what is to be gained from the scoring. Remind participants: "Scoring these assessments may be different from other kinds of scoring you have done. Remember that only certain aspects of the assessments are scored. Other important aspects, for example fit with lessons or amount of student involvement, are not scored because we do not have all the required information."

Explain, sign, and collect confidentiality agreements.

Review the rubrics [use supplementary material as needed, including math and reading standards and Dos and Don'ts for Item Writing]. Stress they should READ the standard, even they think they know it. Read the assessment.

Assign rater codes. Give each rater a number to use as an anonymous identifier. (This is done because data analysis methods required knowing which papers were scored by the same raters.)

Discuss exemplar assessments. Read over each of the anchor assessments and comments. Ask the raters if they can see why each anchor paper received the score it did. Discuss the distinctions between levels. Remind raters to use both the rubrics and the anchor assessments when they do scoring.

Score training assessments.

After reviewing the rubrics and anchors, raters should score a set of 6 training assessments. The goal is to reach at least 80% agreement (70% minimum). Score training assessment A, on each rubric criterion. Raters should record their scores on their score sheets. Then stop, and count the number of raters who assigned each score. Record the tally of how many raters assigned a 1, a 2, and so on, onto the rater agreement summary sheet, as below.

	Criterion 1: Alignment					
Assessment	**1**	**2**	**3**	**4**	**5**	**6**
A	0	0	0	1	8	2
B						
C						
D						
E						
F						

There will be 5 such summary sheets, one for each criterion. One by one, score the next assessments, record scores and score summaries, and discuss. Calculate exact score agreement for each rubric. In the example above, for Alignment for assessment A, rater agreement was 8 out of 11 or 73%. For an overall measure of rater agreement, average the percent agreement across assessments and criteria.

After reliability of scoring is reached, scorers are qualified to assign ratings on their own. However, they may talk with each other if they wish. At times during the scoring session, the facilitator may wish to have a brief discussion with participants. Discussion is fine and will not "taint" scoring. The goal is to have the most reasonable ratings possible for each assessment. Discussion should help, not hinder, that goal.

Study of Local District Assessments
for the
Nebraska Department of Education
Agenda, July 13-14, 2004

Tuesday, July 13, 2004

9:00 – 9:15	Welcome and Introductions Purpose of the Study
9:15 – 10:15	Review of Characteristics of Quality Assessment Review of Rubrics and Exemplars
10:15 – 10:30	Break
10:30 – 12:00	Practice with Rubrics Training to Reliability Criterion and How to Record Data
12:00 – 1:00	Lunch
1:00 - 4:00	Continue Training Session Break as Needed

Wednesday, July 14, 2004

9:00 – 12:00	Scoring Session Break as Needed
12:00 – 1:00	Lunch
1:00 – 3:00	Scoring Session Break as Needed
3:00 – 4:00	Debriefing Session Suggestions for Professional Development in Assessment

Study of Local District Assessments
for the
Nebraska Department of Education
July 13-14, 2004

Purpose of the study. This study fits into a larger program of evaluation of state assessment in Nebraska that includes reviews of district assessment plans, evaluation of the effects of the Nebraska STARS program on schools, teachers, and students, and other components. The research questions for this study are:

- Of what quality are the local assessments used in the Nebraska STARS program? What proportion of them are of sufficient quality to accurately measure student performance?
- What professional development and/or feedback might be helpful to the educators who wrote the assessments?

Purpose for participants in the July 13-14 workshop. You were selected to participate in this workshop because you already have interest and expertise in assessment. We hope that your work here functions as additional professional development in assessment for you. Reviewing a sample of assessments should be a good way to broaden and deepen your own assessment knowledge. We also hope that, based on your review of the particular elements of local assessments, you will be able to suggest specific professional development topics for your colleagues around the state. These two days will include:

- Reviewing the characteristics of quality assessment and training on the use of rubrics to rate that quality. The goal is at least 80% agreement.
- Rating a stratified (by class) random sample of local assessments drawn from district assessment portfolios is Reading and Language Arts (2003) and Mathematics (2004).
- Debriefing the session, to provide: (a) information on what you learned from the session, and (b) what suggestions you have for professional development to meet any needs identified by the review of assessments.

THANK YOU FOR YOUR HELP WITH THIS PROJECT! It is important to the spirit of STARS that Nebraska educators are the reviewers for local assessments.

Study of Local District Asssments
for the Nebraska Department of Education
July 13-14, 2004

CONFIDENTIALITY AGREEMENT

Purpose of the study. This study fits into a larger program of evaluation of state assessment in Nebraska that includes reviews of district assessment plans, evaluation of the effects of the Nebraska STARS program on schools, teachers, and students, and other components. The research questions for this study are:

- Of what quality are the local assessments used in the Nebraska STARS program? What proportion of them are of sufficient quality to accurately measure student performance?
- What professional development and/or feedback might be helpful to the educators who wrote the assessments?

Nature of participation in the July 13-14 workshop. Participants will receive training on a set of rubrics, rate a stratified (by class) random sample of local assessments drawn from district assessment portfolios is Reading and Language Arts (2003) and Mathematics (2004), and participate in discussion.

Risks and benefits: It is expected that this workshop will be a professional development opportunity for participants. It is expected that the Nebraska Department of Education (NDE) will receive useful information about local assessments from the results of this study. No risks to school districts or individuals are expected if participants keep confidential the information they receive as a result of their participation in this workshop. The purpose of this confidentiality agreement is to document participant's agreement to do that.

Confidentiality of information: Researcher/NDE responsibilities. No identification of school district appears on the assessments. Each assessment has been given a unique numerical code and a district code. Any identifying information has been removed from the contents of the assessment. All materials will be stored in a secure location.

Confidentiality of information: Participant responsibilities. Participants agree to keep confidential information they receive as a result of their participation in this workshop. Such information includes: (a) the contents of the assessments themselves; (b) the rubric ratings of assessment quality; and (c) comments made in discussion. Should a participant happen to recognize an assessment, he or she should recuse himself or herself from rating that assessment and not disclose the identity of the district.

I, the undersigned, have read and understood this confidentiality statement and agree to the participant responsibilities.

Name _____

Position _____

Date _____

Nebraska Department of Education
July 13-14, 2004

CLASSROOM ASSESSMENT RATING RUBRIC

ASSESSMENT CODE _____DISTRICT _____ STANDARD _____

RATER _____
Circle the specific part of the criterion that is present and assign points for each criterion.

CRITERIA	1-2 Points	3-4 Points	5-6 Points	SUMMARY POINTS
ALIGNMENT	Few of the assessment items/tasks reflect a match to the appropriate standard.	Some of the assessment items/tasks reflect a match to the appropriate standard(s).	Assessment items/tasks reflect a match to the appropriate standard(s).	
SUFFICIENCY	The essence of the standard is not represented, is ignored, or is poorly sampled.	The essence of the standard is represented by an inadequate number of score points and only at certain performance levels.	The essence of the standard is represented by an appropriate number of score points at all four performance levels (required by 2006-07).	
CLARITY	Few of the assessment tasks and/or directions are clear, complete, or unambiguous.	Some of the assessment tasks or the directions are clear, complete, or unambiguous.	The assessment tasks and the directions are clear, complete, and unambiguous.	
APPROPRIATENESS	Few of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length.	Some of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length.	Assessment tasks are fair, at the appropriate cognitive level, and of an appropriate length.	
SCORING PROCEDURES	Scoring procedures/rubrics are inadequate or not provided.	Scoring procedures/ rubrics are provided, but require judgments that would be hard to make fairly and consistently.	Scoring procedures/ rubrics are consistent with the assessment task and can be applied clearly and consistently.	

Study of Local District Assessments for the Nebraska Department of Education
July 13-14, 2004

DEBRIEFING QUESTIONS Reading___ Math___

Please reflect on your experiences these past two days: learning about (or reviewing) characteristics of quality assessment, learning how to use the rubrics reliably, and reading and scoring assessments used by your colleagues around the state.

1. What were your major learnings from this experience? [This could be related to the standards, the rubrics, or any of the resource material; it could also be related to the opportunity to see so many examples of different assessments from around the state.] Please be as specific as you can.

2. Did you notice any patterns in the assessments that you scored that might suggest particular assessment topics for professional development?

3. What kinds of feedback do you think might be helpful to the educators who crafted the assessments you scored?

4. Do you think you will use any of what you learned here back at your own school?

   a. What, if anything, that you learned might help you in your own work?

   b. What, if anything, that you learned here might help your colleagues?

   c. What, if anything, that you learned here might help your students?

RATER AGREEMENT CALCULATIONS

CRITERION 1: ALIGNMENT						
Assessment	**1**	**2**	**3**	**4**	**5**	**6**
A						
B						
C						
D						
E						
F						

CRITERION 2: SUFFICIENCY						
Assessment	**1**	**2**	**3**	**4**	**5**	**6**
A						
B						
C						
D						
E						
F						

CRITERION 3: CLARITY						
Assessment	**1**	**2**	**3**	**4**	**5**	**6**
A						
B						
C						
D						
E						
F						

CRITERION 4: APPROPRIATENESS						
Assessment	**1**	**2**	**3**	**4**	**5**	**6**
A						
B						

C						
D						
E						
F						

CRITERION 5: SCORING PROCEDURES						
Assessment	1	2	3	4	5	6
A						
B						
C						
D						
E						
F						

Local District Assessments:
One Indicator for the Validation of Nebraska's
Standards, Assessment, and Accountability System

Appendix B

Reliability Results

Reliability of Scores from the
Nebraska Classroom Assessment Rating Rubric
July 13-14, 2004

Rater Training

The first day of the scoring workshop, July 13, was devoted to rater training. Participants discussed four example assessments, two math and two reading. These examples had been selected to demonstrate a range of quality and had already been rated (by the researcher and an NDE assessment staff member, with 100% agreement). The rubric sheets had been filled out for these assessments and were attached to each as a cover sheet. The group discussed each example in turn, including why it had received the ratings it did.

After discussing the example assessments, all 13 raters then rated a training assessment independently, using the rubric sheet. The facilitator asked, and recorded, what quality rating each individual had assigned for each of the 5 criteria. Discussion followed, but was not allowed to change the independent ratings. A second training assessment was then scored independently, its ratings were recorded, and discussion followed. In all, five training assessments were completed using this process.

At the end of the day, rater agreement overall (13 raters rating 5 training assessments on 5 criteria) was 86% exact agreement on category (low/medium/high) and 66% exact agreement on point value (1-6). Percent agreement had stabilized at this level after the 4th training assessment. The second day, raters scored the remaining assessments independently.

Double Scoring Study

On July 14, raters selected assessments from six folders (Math and Reading, grades 4, 8, and 11, respectively). Raters were deemed "reliable" for any of the rating (the agreement figures had been on a mixture of subject areas) but were encouraged to rate assessments of subjects and levels where they felt most expert. Five of the assessments (randomly selected) in each of the folders had been included twice, and thus were scored by two raters. Raters who pulled a copy of an assessment they already scored were instructed to return it to the folder for someone else to select.

Two kinds of analysis were possible for the double scoring study. First, for the entire set of 30, overall and by subject, intraclass correlations were calculated for each rubric criterion separately and for total score. These results (Table B.1 below) indicated that the math ratings remained more reliable than the reading ratings. The most reliable ratings were for Clarity and Scoring Procedures.

Second, because the assessments had been sorted into grade and subject level and raters had selected assessments on that basis, there were 5 subsets (3 math and 2 reading) where the same two raters had rated the same 3-6 assessments. In all, these involved 20 of the 30 assessments in the double scoring study. For these subsets, generalizability studies were performed. The results

of this analysis (Table B.2 below) also indicated that the math ratings remained more reliable than the reading ratings.

Discussion

Based on the training reliability, expressed as percent agreement, rubric judgments of raters during the following days are well worth attending to, and do indicate the general level of quality of all 296 of the assessments in the final sample. Had the study design not included an embedded double scoring study, the percent agreement after training would have been sufficient to warrant the main study ratings.

However, the results of the double scoring study indicate that it was more difficult for Reading raters than for Math raters to maintain their level of reliability. The results from this first rollout of the rubrics give indications of where additional work on reliability might be targeted. The positive reaction of the panelists (see full report) indicated that they viewed the rubrics as potentially very reliable and even prescriptive for developing and self-assessing the quality of local assessments in districts.

In particular, Sufficiency seemed difficult for Reading raters to judge reliably. Appropriateness seemed difficult for both Math and Reading raters to judge reliably. Not surprisingly, these are two areas that are under current discussion in the Nebraska STARS system, and it is logical that there would be more difficulty with these areas.

Regarding Alignment, educators have been working with the Nebraska Standards for years and are familiar with them, and familiar with the idea of using them as the basis for assessments. Regarding Clarity (to students), this is a focus of teachers and educators at every level. Observation and discussion during the scoring workshop indicated that judging whether the assessments would be clear to students was the judgment raters found easiest to make. Regarding Scoring Procedures, a large part of the judgment was simply that no scoring procedures were provided. In fact, at the beginning of the training the Scoring Procedures rubric was going to be counted in the top category if all items required right/wrong scoring and the only thing missing was an answer key (districts simply did not realize an answer key was considered part of the assessment and did not include it, where it must have existed). However, the raters found that judging "required right/wrong scoring and the only thing missing was an answer key" was something they could not agree on, and for reliability the group agreed to make a literal "not provided" judgment if scoring information was not provided, no matter how straightforward it might have been.

Recommendations

Recommendations to raise the reliability of the use of the rubrics include:

- continued practice with the rubrics, including use of more examples (recommended by participants during the debriefing as a method for when they use the rubrics with districts during the academic year);

- slight revision of the language in the rubrics to eliminate some confusion and increase clarity (*e.g*., remove "essence of the standard" from the Sufficiency descriptions; see full report for complete recommendations for revisions); and

- continued discussions, especially among Reading educators, about the scope of the standards.

Table B.1. Intraclass Correlations (For Single Rater, One Way Random Effects Model)						
	Math		Reading		Overall	
Score	**ICC**	**95% CI**	**ICC**	**95% CI**	**ICC**	**95% CI**
Alignment	.28	-.22 to .67	.22	-.32 to .65	.27	-.09 to .57
Sufficiency	.40	-.08 to .74	.08	-.44 to .57	.20	-.15 to .52
Clarity	.64	.25 to .86	.60	.14 to .85	.61	.34 to .80
Appropriateness	.16	-.34 to .59	-.25	-.67 to .29	-.15	-.48 to .21
Scoring Procedure	.61	.19 to .84	.26	-.27 to .68	.48	.15 to .71
Total Score	.59	.16 to .83	-.13	-.59 to .41	.10	-.26 to .44

ICC – Intraclass correlation
95% CI – 95% Confidence Interval

Table B.2. Generalizability Analysis Decision Study for 1 Rater (raters random), 5 Items (items fixed)					
Data subset (total 20 different assessments)	Subject	Generalizability for Relative Decisions		Generalizabilty for Absolute Decisions	
		$E\hat{\rho}^2$	Signal-noise Ratio	Phi	Signal-noise Ratio
6 assessments	Math	.78	3.64	.65	1.85
3 assessments	Math	.81	4.29	.81	4.29
4 assessments	Math	.90	9.49	.88	7.38
3 assessments	Reading	.00	.00	.00	.00
4 assessments	Reading	.00	.00	.00	.00

Local District Assessments:
One Indicator for the Validation of Nebraska's
Standards, Assessment, and Accountability System

Appendix C

Number and Percent of Assessments at
Different Levels of Quality

By Grade and Subject

MATHEMATICS, GRADE 4 Alignment (Total = 49 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Assessment items/tasks reflect a match to the appropriate standard(s).	6	39	79.6
	5	4	8.2
Some of the assessment items/tasks reflect a match to the appropriate standard(s).	4	6	12.2
	3	0	0.0
Few of the assessment items/tasks reflect a match to the appropriate standard.	2	0	0.0
	1	0	0.0

MATHEMATICS, GRADE 4 Sufficiency (Total = 49 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
The essence of the standard is represented by an appropriate number of score points at all four performance levels (required by 2006-07).	6	16	32.7
	5	15	30.6
The essence of the standard is represented by an inadequate number of score points and only at certain performance levels.	4	18	36.7
	3	0	0.0
The essence of the standard is not represented, is ignored, or is poorly sampled.	2	0	0.0
	1	0	0.0

MATHEMATICS, GRADE 4 Clarity (Total = 49 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
The assessment tasks and the directions are clear, complete, and unambiguous.	6	36	73.5
	5	6	12.2
Some of the assessment tasks or the directions are clear, complete, or unambiguous.	4	7	14.3
	3	0	0.0
Few of the assessment tasks and/or directions are clear, complete, or unambiguous.	2	0	0.0
	1	0	0.0

MATHEMATICS, GRADE 4 Appropriateness (Total = 49 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Assessment tasks are fair, at the appropriate cognitive level, and of an appropriate length.	6	28	57.1
	5	15	30.6
Some of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length.	4	5	10.2
	3	1	2.0
Few of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length.	2	0	0.0
	1	0	0.0

MATHEMATICS, GRADE 4 Scoring Procedures (Total = 49 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Scoring procedures/ rubrics are consistent with the assessment task and can be applied clearly and consistently.	6	11	22.4
	5	3	6.1
Scoring procedures/ rubrics are provided, but require judgments that would be hard to make fairly and consistently.	4	2	4.1
	3	3	6.1
Scoring procedures/rubrics are inadequate or not provided.	2	26	53.1
	1	4	8.2

MATHEMATICS, GRADE 8 Alignment (Total = 50 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Assessment items/tasks reflect a match to the appropriate standard(s).	6	32	64.0
	5	9	18.0
Some of the assessment items/tasks reflect a match to the appropriate standard(s).	4	7	14.0
	3	2	4.0
Few of the assessment items/tasks reflect a match to the appropriate standard.	2	0	0.0
	1	0	0.0

MATHEMATICS, GRADE 8 Sufficiency (Total = 50 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
The essence of the standard is represented by an appropriate number of score points at all four performance levels (required by 2006-07).	6	5	10.0
	5	23	46.0
The essence of the standard is represented by an inadequate number of score points and only at certain performance levels.	4	19	38.0
	3	2	4.0
The essence of the standard is not represented, is ignored, or is poorly sampled.	2	1	2.0
	1	0	0.0

MATHEMATICS, GRADE 8 Clarity (Total = 49 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
The assessment tasks and the directions are clear, complete, and unambiguous.	6	42	85.7
	5	6	12.2
Some of the assessment tasks or the directions are clear, complete, or unambiguous.	4	1	2.0
	3	0	0.0
Few of the assessment tasks and/or directions are clear, complete, or unambiguous.	2	0	0.0
	1	0	0.0

MATHEMATICS, GRADE 8 Appropriateness (Total = 50 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Assessment tasks are fair, at the appropriate cognitive level, and of an appropriate length.	6	25	50.0
	5	15	30.0
Some of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length.	4	7	14.0
	3	3	6.0
Few of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length.	2	0	0.0
	1	0	0.0

MATHEMATICS, GRADE 8 Scoring Procedures (Total = 50 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Scoring procedures/ rubrics are consistent with the assessment task and can be applied clearly and consistently.	6	10	20.0
	5	2	4.0
Scoring procedures/ rubrics are provided, but require judgments that would be hard to make fairly and consistently.	4	5	10.0
	3	2	4.0
Scoring procedures/rubrics are inadequate or not provided.	2	27	54.0
	1	4	8.0

MATHEMATICS, GRADE 11 Alignment (Total = 48 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Assessment items/tasks reflect a match to the appropriate standard(s).	6	34	70.8
	5	2	4.2
Some of the assessment items/tasks reflect a match to the appropriate standard(s).	4	9	18.8
	3	0	0.0
Few of the assessment items/tasks reflect a match to the appropriate standard.	2	3	6.3
	1	0	0.0

MATHEMATICS, GRADE 11 Sufficiency (Total = 48 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
The essence of the standard is represented by an appropriate number of score points at all four performance levels (required by 2006-07).	6	10	20.8
	5	18	37.5
The essence of the standard is represented by an inadequate number of score points and only at certain performance levels.	4	16	33.3
	3	1	2.1
The essence of the standard is not represented, is ignored, or is poorly sampled.	2	3	6.3
	1	0	0.0

MATHEMATICS, GRADE 11 Clarity (Total = 48 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
The assessment tasks and the directions are clear, complete, and unambiguous.	6	32	66.7
	5	6	12.5
Some of the assessment tasks or the directions are clear, complete, or unambiguous.	4	9	18.8
	3	1	2.1
Few of the assessment tasks and/or directions are clear, complete, or unambiguous.	2	0	0.0
	1	0	0.0

MATHEMATICS, GRADE 11 Appropriateness (Total = 48 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Assessment tasks are fair, at the appropriate cognitive level, and of an appropriate length.	6	36	75.0
	5	1	2.1
Some of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length.	4	10	20.8
	3	0	0.0
Few of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length.	2	1	2.1
	1	0	0.0

MATHEMATICS, GRADE 11 Scoring Procedures (Total = 48 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Scoring procedures/ rubrics are consistent with the assessment task and can be applied clearly and consistently.	6	5	10.4
	5	0	0.0
Scoring procedures/ rubrics are provided, but require judgments that would be hard to make fairly and consistently.	4	1	2.1
	3	0	0.0
Scoring procedures/rubrics are inadequate or not provided.	2	38	79.2
	1	4	8.3

READING, GRADE 4 Alignment (Total = 50 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Assessment items/tasks reflect a match to the appropriate standard(s).	6	27	54.0
	5	9	18.0
Some of the assessment items/tasks reflect a match to the appropriate standard(s).	4	5	10.0
	3	5	10.0
Few of the assessment items/tasks reflect a match to the appropriate standard.	2	2	4.0
	1	2	4.0

READING, GRADE 4 Sufficiency (Total = 50 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
The essence of the standard is represented by an appropriate number of score points at all four performance levels (required by 2006-07).	6	14	28.0
	5	11	22.0
The essence of the standard is represented by an inadequate number of score points and only at certain performance levels.	4	15	30.0
	3	1	2.0
The essence of the standard is not represented, is ignored, or is poorly sampled.	2	8	16.0
	1	1	2.0

READING, GRADE 4 Clarity (Total = 50 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
The assessment tasks and the directions are clear, complete, and unambiguous.	6	36	72.0
	5	5	10.0
Some of the assessment tasks or the directions are clear, complete, or unambiguous.	4	2	4.0
	3	0	0.0
Few of the assessment tasks and/or directions are clear, complete, or unambiguous.	2	6	12.0
	1	1	2.0

READING, GRADE 4 Appropriateness (Total = 50 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Assessment tasks are fair, at the appropriate cognitive level, and of an appropriate length.	6	21	42.0
	5	15	30.0
Some of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length.	4	10	20.0
	3	0	0.0
Few of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length.	2	3	6.0
	1	1	2.0

READING, GRADE 4 Scoring Procedures (Total = 50 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Scoring procedures/ rubrics are consistent with the assessment task and can be applied clearly and consistently.	6	11	22.0
	5	5	10.0
Scoring procedures/ rubrics are provided, but require judgments that would be hard to make fairly and consistently.	4	11	22.0
	3	3	6.0
Scoring procedures/rubrics are inadequate or not provided.	2	18	36.0
	1	2	4.0

READING, GRADE 8 Alignment (Total = 46 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Assessment items/tasks reflect a match to the appropriate standard(s).	6	24	52.2
	5	7	15.2
Some of the assessment items/tasks reflect a match to the appropriate standard(s).	4	12	26.1
	3	3	6.5
Few of the assessment items/tasks reflect a match to the appropriate standard.	2	0	0.0
	1	0	0.0

READING, GRADE 8 Sufficiency (Total = 46 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
The essence of the standard is represented by an appropriate number of score points at all four performance levels (required by 2006-07).	6	17	37.0
	5	8	17.4
The essence of the standard is represented by an inadequate number of score points and only at certain performance levels.	4	13	28.3
	3	1	2.2
The essence of the standard is not represented, is ignored, or is poorly sampled.	2	6	13.0
	1	1	2.2

READING, GRADE 8 Clarity (Total = 45 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
The assessment tasks and the directions are clear, complete, and unambiguous.	6	22	48.9
	5	5	11.1
Some of the assessment tasks or the directions are clear, complete, or unambiguous.	4	8	17.8
	3	2	4.4
Few of the assessment tasks and/or directions are clear, complete, or unambiguous.	2	7	15.6
	1	1	2.2

READING, GRADE 8 Appropriateness (Total = 46 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Assessment tasks are fair, at the appropriate cognitive level, and of an appropriate length.	6	23	50.0
	5	7	15.2
Some of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length.	4	13	28.3
	3	2	4.3
Few of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length.	2	1	2.2
	1	0	0.0

READING, GRADE 8 Scoring Procedures (Total = 46 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Scoring procedures/ rubrics are consistent with the assessment task and can be applied clearly and consistently.	6	11	23.9
	5	3	6.5
Scoring procedures/ rubrics are provided, but require judgments that would be hard to make fairly and consistently.	4	7	15.2
	3	1	2.2
Scoring procedures/rubrics are inadequate or not provided.	2	15	32.6
	1	9	19.6

READING, GRADE 11 Alignment (Total = 50 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Assessment items/tasks reflect a match to the appropriate standard(s).	6	27	54.0
	5	3	6.0
Some of the assessment items/tasks reflect a match to the appropriate standard(s).	4	15	30.0
	3	0	0.0
Few of the assessment items/tasks reflect a match to the appropriate standard.	2	3	6.0
	1	2	4.0

READING, GRADE 11 Sufficiency (Total = 50 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
The essence of the standard is represented by an appropriate number of score points at all four performance levels (required by 2006-07).	6	15	30.0
	5	7	14.0
The essence of the standard is represented by an inadequate number of score points and only at certain performance levels.	4	20	40.0
	3	1	2.0
The essence of the standard is not represented, is ignored, or is poorly sampled.	2	5	10.0
	1	2	4.0

READING, GRADE 11 Clarity (Total = 50 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
The assessment tasks and the directions are clear, complete, and unambiguous.	6	29	58.0
	5	7	14.0
Some of the assessment tasks or the directions are clear, complete, or unambiguous.	4	9	18.0
	3	0	0.0
Few of the assessment tasks and/or directions are clear, complete, or unambiguous.	2	5	10.0
	1	0	0.0

READING, GRADE 11 Appropriateness (Total = 50 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Assessment tasks are fair, at the appropriate cognitive level, and of an appropriate length.	6	33	66.0
	5	11	22.0
Some of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length.	4	5	10.0
	3	0	0.0
Few of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length.	2	1	2.0
	1	0	0.0

READING, GRADE 11 Scoring Procedures (Total = 50 assessments)			
Quality Level Description	**Points**	**Number of Assessments**	**Percent of Assessments**
Scoring procedures/ rubrics are consistent with the assessment task and can be applied clearly and consistently.	6	10	20.0
	5	4	8.0
Scoring procedures/ rubrics are provided, but require judgments that would be hard to make fairly and consistently.	4	10	20.0
	3	0	0.0
Scoring procedures/rubrics are inadequate or not provided.	2	24	48.0
	1	2	4.0

Local District Assessments:
One Indicator for the Validation of Nebraska's
Standards, Assessment, and Accountability System

Appendix D

Verbatim Comments
from Debriefing Questionnaires

and

Facilitator Notes
from Debriefing Session

DEBRIEFING QUESTIONS
Verbatim responses from participant surveys, N=13, 100%

1. What were your major learnings from this experience?

1-R
- Seeing other assessments helped me put ours in perspective
- Working w/ this rubric at this level of development

2-RM
~ great seeing examples from around state - rubric was helpful…will be great too!

3-RM
Understanding of the rubric. Ways to change the rubric to make it clearer.

4-RM
- Learning of a systematic/ consistent process to rate/judge assessments
- Learning the use of the rubric, in order to train others
- Clarity of assessments

5-R
The #1 item I learned is how to better evaluate if an assessment matches the essence of the standard.

6-M
- Using a rating rubric to evaluate the classroom assessment rating.
- Looking at the variety of assessments

7-RM
I have a much clearer picture of what a quality assessment looks like and what specifically needs to be present to make it a quality measuring tool. It was extremely beneficial to see assessment copies ranging from 1 to 6 points. I am anxious to look at my home district assessments. I am extremely grateful for this training. This is a piece that was missing and I think it will help districts become more accountable.

8-M
Assessments can really only be a sampling of the standards. A test can never be all inclusive. Sufficiency is hard to insure.

9-RM
I'm delighted to actually look at assessments and to have some criteria for judging "goods" and "bads."

10-R
My major learning is that we have much work to do in helping teachers (I read only grade 11 assessments) understand how to make better connections between their assessments and the standards.

11-M

Learning the criteria for the classroom (rubric). The conversation in the specificity of the criteria was very helpful. The more content specific a person is to review an assessment – the more accurate the analysis. A process to use as well.

12-R
I liked learning specific ways to evaluate teacher made assessments.

13-[subject not identified]
- This experience validated that much assessment knowledge has been gained and applied in our state.
- While there are many quality assessments currently being used to assess student performance, some assessment revision is needed.
- The department of education is committed to furthering assessment literacy and to the STARS process. It is obvious that constant planning occurs to ensure that our assessment process continues improving.

2. Did you notice any patterns in the assessments that you scored that might suggest particular assessment topics for professional development?

1-R
- We must provide a comprehensive training on using this rubric with samples across the state!
- Also teachers need to understand that all standards are not "equal" in cognitive requirements for application/processing. Some are addressing skills and some require knowledge and others concepts (application). They are not seeing that.

2-RM
high school teachers -> assessments were poorly written, disappointing … bring h.s. teachers together for workshop & 'redo'/revise assessments in large groups (HS teachers are always smallest group)

3-RM
- Scoring procedures were missing, but I think it's unfair to say districts don't have them because the Dept. has never required them. Clarity was good on almost all of them.

4-RM
- Scoring/directions – weak points. This really is an area that needs work.
- Sufficiency is also an area of concern, demonstrated by a pattern of lower ratings.

5-R
Performance based assessments need teaching units developed related to them & included in the assessment package.

6-M
- Scoring procedures/rubrics need to be included in portfolio.
- Subjective items required a rubric.

7-RM

Those with local standards did not seem equal or more rigorous than! Scoring procedures were obviously an issue. I do think it would help districts to see high quality scoring guidelines. I also think districts need to re-examine how well their assessments cover the entire standard and not just portions of the standard. Sufficiency is an issue but again not required until 2006/07.

8-M
Sufficiency – include scoring guides. Local standards never seem to be as stringent as the state standards.

9-RM
Remind people to be specific about scoring, points allowed, items weighted, partial credit, spelling counts, complete sentences required, anything.

10-R
My lowest ratings occurred in sufficiency and scoring procedures.

11-M
- Scoring procedures were seldom provided.
- Grade level appropriateness was questioned.
- All 4 proficiency performance levels were not present.

12-R
Standards were interpreted within quite a range for alignment. I think the verbs used in the standards have caused some confusion, especially the verb "identify." ("Identify" is often met by the verb that follows it.)

13-[subject not identified]
- Alignment was an issue with numerous assessments in that on some assessments the standard was not adequately represented (content/sufficiency).

3. What kinds of feedback do you think might be helpful to the educators who crafted the assessments you scored?

1-R
- Better language on rubric still

2-RM
- Why we scored assessments low
- suggestions to make it better
- examples of good assessments

3-RM
District personnel need to take the time to look at their own assessments. It would be helpful for districts to get a "full" training on the rubric and then do their work locally.

4-RM

- Evaluate organization & completeness of assessments
- Provide cover sheet w/ appropriate scoring directions
- Revisit sufficiency levels

5-R
The chats should work excellently so educators can evaluate their own assessments.

6-M
- Information on match – other types of items used to match standards – different from theirs.
- Item clarity or sufficiency for assessments – how to improve.
- Specific feedback on scoring process and how to use data.
-

7-RM
The rubric would be great feedback. Any conversation that a rater could have would also be helpful. I wonder if there are trends in the scores? Does a district generally have high scores or low scores in the same area across the board?

8-M
The assessments almost always seemed to reflect the standards. We (in NE) are good at the fit. The sufficiency aspect needs to be developed.

9-RM
It would be great if they could do what we did. They'd see the value in a complete cover letter, scoring requirements, and sufficiencies.

10-R
I think it is imperative that these assessment writers look carefully at how and why they decide what is important to assess.

11-M
- Description of criteria
- How to improve local assessments
- Excellent [sic … example?] of a good, middle & poor assessment

12-R
I think the rubric will be helpful to schools.

13-[subject not identified]
- Keep revising and seeking improved documents.
- Network with other districts!

4. Do you think you will use any of what you learned here back at your own school?

[13 yes, 100%]

a. What, if anything, that you learned might help you in your own work?

1-R
All of it! I've been wanting a rubric like this & feel that this will be very helpful.

2-RM
rubric – part of assessment writing process

3-RM
Actually working one-on-one w/ districts to do the process

4-RM
Offering inservice training to local review teams & teacher teams.

5-R
We will definitely evaluate our assessments using the rubric.

6-M
Using a rubric to evaluate assessment.

7-RM
We will be revising our LA & Math assessments. Now I have a better idea of what we need to look for. It has been extremely helpful to see sample assessments from other districts – there are some great examples out there. It would be nice to have a way to see others to get the ideas. We create these in isolation and seeing others work would be helpful!

8-M
We will use the rubric. I will include answer sheets in the portfolio.

9-RM
I intend to run our assessment/curriculum committee through some samples & work/talk together.

10-R
All of my learnings over these two days about the qualities of assessments that we deem important will help me in my work w/ schools, ESUs, and NDE colleagues.

11-M
Classroom assessment rubric will be used in the development of our classroom assessments for the SI goal.

12-R
This process cleared up expectations.

13-[subject not identified]
I will look for the revised rubric in order to do local assessment review.

b. What, if anything, that you learned might help your colleagues?

1-R
All of it! As a staff dev. I see that we need this training so very badly!

2-RM
rubric good/poor examples would be great to share

3-RM
Actually working one-on-one w/ districts to do the process

4-RM
Using the process will obviously aid in the development of better assessments

5-R
How to evaluate assessments in order to better develop quality assessments.

6-M
Same – science & social studies evaluation.

7-RM
clarify scoring directions for teachers – include sample responses to open ended or essays – clarify rubrics

8-M
The test can never be all inclusive, rather it is simply a sampling.

9-RM
After working with the curriculum committee, they'll share the "meat" of the material with their colleagues.

10-R
All of my learnings over these two days about the qualities of assessments that we deem important will help me in my work w/ schools, ESUs, and NDE colleagues.

11-M
Classroom assessment rubric will be used in the development of our classroom assessments for the SI goal.

12-R
This is a concrete process that would help schools review their assessments in an objective manner.

13-[subject not identified]
Assessing assessment can teach much about assessment development/revision.

c. What, if anything, that you learned might help your students?

1-R

All of it! This process is critical for students to be assessed fairly.

2-RM
make assessments more clear/'friendly' for student … stop overassessing.

3-RM
N/A

4-RM
The better the assessment, the more accurate the student results should be

5-R
Student learning goes up when assessments are high quality because teaching becomes more directed.

6-M
Better assessments that measure the intended standards and provide appropriate & valid feedback to student.

7-RM
The need to clarify directions for the student (and the teacher)

8-M
Once the target is identified, the students can know where they are going.

9-RM
(blank)

10-R
(blank)

11-M
- The criteria "clarity."
- The need for scoring procedures may be helpful as well.
- The quality of assessments will lead to higher quality instruction = increased student achievement.

12-R
Better assessments improve the learning process.

13-[subject not identified]
Quality assessment will help ensure valid assessment results.

DEBRIEFING DISCUSSION
Facilitator notes from group discussion (13 participants), July 14, 2004

What did you learn from participating in this scoring workshop?

- Districts should not get individual feedback because a lot of the scoring procedures were not provided (but did exist).
- This process to use with review teams and staff is clearer than coming up with one on our own.
- In Nebraska we have to "live it" for a while.
- We've just broken our #1 rule – they (portfolio preparers) never knew what the assessment was.
- This is a good snapshot for a district of another piece (of the assessment system) that was missing before.
- Like when you pilot a writing prompt and then pull anchors – in a sense that's what we were doing – it's very hard, but we couldn't have done it (reviewed assessments) before.
- Shock – disconnect between portfolio scoring – assuming they're doing the work with high quality assessments – they may not be (e.g. high school Reading, not unpacking the standards and not understanding good assessment).
- Worry that we're leaning toward simplistic assessments – maybe encouraged by the number of times the standards use "identify."
- There are a lot of quality assessments, too, that are fun to look at.
- Aurora decided to look at assessments; we validated how far we've come but we have lots of educating yet to do.
- Among new educators, there are only a small group of high school teachers. There are not many (h.s. teachers) in representative groups like you have in elementary. There are not as many high school people represented.
- People at the ground level are working; it's a growth process.
- If assessments were poorly constructed because teachers didn't work very hard, that's scary. But it's even scarier if that (poor sample on the assessment) was what teachers thought was important for students to know from those standards.
- Maybe there are still some singletons.

- One needs to know how to do something – process (portfolio) first, product (assessment) next. Some clearly understood the standards. Now we need to take a broader look (big picture) at how the standards fit into the bigger picture.

- Now that we're starting to put criteria on assessments, teachers can grow; we can all grow.

- We need to build our pedagogical vocabulary in assessment.

- In performance based assessment, the assignments need to be there. (Without that) it's almost like writing assessments without the prompt.

- A potential criterion (for assessment) is focus – focus needs to be embedded in day to day instruction until they've connected assessment with instruction, and it's not just an add-on.

- Assessment cannot wait until the end of the semester. Students need to be assessed in ongoing instruction.

- NDE has the School Improvement Process, and there are differences between those who think that's a "layer" and ones who think school improvement is part of their regular work. NDE can tell from talking with districts, for the ones who are getting unacceptable ratings, it's a "layer."

- Student self-assessment is important. Teachers who are assessment literate know how important that is. But that's not common knowledge. We can continue to build this process. When we share this rubric, we should add that we think you should build in a student self-assessment concept. Some standards lend themselves to it.

- This process validates that NDE is interested in increasing assessment literacy.

- The goal of assessment review visits was to look at assessments, but they turned into portfolio sessions.

- The state needs information on match to standard. There is a misconception (among districts) that the assessment was reviewed and assessed for quality (in the portfolio rating process).

- The six quality criteria for portfolios did start the process of raising assessment literacy.

- "Had I just seen some of that" ours (district assessment) would have been so much better. For example, what students need to know to do this assessment. Examples (assessments) help (illustrate better assessment), similar to sample portfolios (helping with portfolios).

- Maybe pull examples of five's or six's (assessments that scored 5-6 on the rubric).

- Sometimes the ratings wouldn't tell you what to fix. Recommend having a comments space on the bottom (of the rubric) to tell what to fix.

- Need a description of what makes quality.

- Could have some samples of good and poor assessments, make sure to say it's just one way (not to just copy sample assessments). "Adapt" vs. "adopt" – if you show an exemplary assessment lots of people will adopt it.

- For some of the assessments of local standards, they got 5's or 6's but the standard said something like "will read literature in English" – serious problem with "equal to or more rigorous than." In law there's a term: "void for vagueness." I feel like some of the local standards are void for vagueness.

- We're ready and seeing a need to go back to the assessment development level. Why not revisit the standards – Nebraska as well as local.

- We need to keep our work at a practical level. A need for doing "more things" may overwhelm local educators.

- Keep in mind the purposes of this assessment – accountability, not diagnostic.

- How can we continue to talk about assessment in a vacuum? It needs to be tied to instruction.

- We would need too much information for (assessments to be useful for) diagnostic purposes. It would be too massive for accountability purposes. Classroom assessments need to do that (give diagnostic information).

- For Nebraska approved local standards, Nebraska says they'll honor local assessments.

- Local scoring events yield good discussions.

- Sample assessments might be used well, not just copied. We need to show examples, and lots of them.

Based on what you saw during the scoring, what professional development needs do you see?

- Explicit scoring procedures

- Sufficiency – performance is not represented at the 4 levels. (Many of the assessments are) Mastery/non-Mastery tests. There's very little at the low or high end. Lots of the same kinds of questions, with levels defined as the number correct, without regard to the level of difficulty of the question.

- Alignment – need a description of what teachers thought the standard was. They need professional development on (1) item writing to standards that are understood, and (2) showing they understood what the standards was to begin with. (Example – understanding of "organization" in giving oral presentations vs. "organization" of content in the speech.) Sometimes they looked at verbs (in the standard), sometimes they need to look at the content.

- Example indicators (in the standards) are optional or are only examples…they're not all exceptionally good examples or necessarily "capturing the essence" of the standard.

- Knowledge base (of the teachers constructing the assessments) needs to be as broad as possible.

- Formatting is still an issue with some assessments: spelling, design of answer spaces, etc. Need professional development on good assessment design practices. From a kid's perspective, it needs to be clear what to "do" (how to respond to answer the question). Fitting responses into little spaces, etc. … Formatting would make an easy inservice.

- Need to clarify the difference between sufficiency and appropriateness (having questions at a range of difficulty on the standard but still being "grade level appropriate").

- Need professional development on the criteria for developmental appropriateness (vocabulary, etc.) over a range of "meatiness" of standards. In a district, elementary and high school level of concepts – use student data to support these suggestions – involve students in feedback about assessments.

- Will teachers be critical of their own assessments? Need administrator support.

- Need to use classroom assessments for the school improvement process goals. Predict CRT and NRT (from classroom assessments).

- Rubric could be expanded. If we have a poor assessment, how do we know it and how do we improve it?

- Getting together with people, learn from each other.